Supplemental Material

CBE—Life Sciences Education

Dasgupta et al.

SUPPLEMENTARY INFORMATION

THE ‘SHRIMP ASSESSMENT’

Assessment: The College Board (2006) AP® Statistics Free-Response Question 5 Page 9. [Online http://apcentral.collegeboard.com/apc/public/repository/_ap06_frq_statistics_51653.pdf]
Scoring Guidelines:
http://apcentral.collegeboard.com/apc/public/repository/_ap06_statistics_sg_revised.pdf (Page 16)
(Used with permission to Nancy Pelaez, npelaez@purdue.edu)

Background Information
A biologist is interested in studying the effect of growth-enhancing nutrients and different salinity (salt) levels in water on the growth of shrimps. The biologist has ordered a large shipment of young tiger shrimps from a supply house for use in the study. The experiment is to be conducted in a laboratory where 10 tiger shrimps are placed randomly into each of 12 similar tanks in a controlled environment. The biologist is planning to use 3 different growth-enhancing nutrients (A. B. and C) and two different salinity levels (low and high).

1. List the treatments that the biologist plans to use in this experiment.
The three different growth-enhancing nutrients (A, B, and C) and two different salinity levels (low and high) yield a total of 3*2 = 6 different treatment combinations for this experiment, so each can be replicated.

Treatment	Salinity	Nutrient
1	Low	A
2	High	A
3	Low	B
4	High	B
5	Low	C
6	High	C

2. Using the treatments listed in part (a), describe a completely randomized design that will allow the biologist to compare the shrimps' growth after 3 weeks.
Since 10 tiger shrimps have already been randomly placed into each of 12 similar tanks in a controlled environment, we must randomly assign the treatment combinations to the tanks. Each treatment combination will be randomly assigned to 2 of the 12 tanks. One way to do this is to generate a random number for each tank. The treatment combinations are then assigned by sorting the random numbers from smallest to largest.

3. Give one statistical advantage to having only tiger shrimps in the experiment. Explain why this is an advantage.
Using only tiger shrimp will reduce a source of variation in the experimental units, the tanks of shrimp in this experiment. By eliminating this possible source of variation, type of shrimp, we are better able to isolate the variability due to the factors of interest to us (nutrient and salinity level). This will make it easier to identify any treatment effects that may be present.

4. Give one statistical disadvantage to having only tiger shrimps in the experiment. Explain why this is a disadvantage.
Using only tiger shrimp will limit the scope of inference for the biologist. Ideally, the biologist would like to identify the treatment combination that leads to the most growth for all shrimp. However, the biologist will only be able to identify the best treatment combination for tiger shrimp because other types of shrimp may respond differently to the treatments.

THE 'DRUG ASSESSMENT'

Assessment: [©1997-2005 SRI International, Center for Technology in Learning. All rights reserved. http://pals.sri.com/tasks/9-12/Testdrug/]
Scoring Guidelines: http://pals.sri.com/tasks/9-12/Testdrug/rubric.html
Contributed by: New York State Alternative Assessment in Science Project (NYSED)]

Background

The drug ALAMAIN has been developed by the Gentronic Drug Company to lower blood pressure in people whose blood pressure is too high. The drug has been thoroughly tested on animals with positive results. The Gentronic Drug Company feels it is now time for the drug to be tested on humans, and have contacted the Human Improvement Laboratory (HIL) to do the testing.

Directions

As chief research scientist at the Human Improvement Laboratory (HIL) you have been assigned the task of developing the human testing program for the new high blood pressure drug Alamain. You and your assistants are to confer on the experimental design of this testing program, and to write a report outlining the program. The report is to be submitted to the chairperson of the HIL Drug-Testing Committee for approval. Complete the following sections as you would include them on your report.

1. Using complete sentences state the hypothesis to be tested.
Alamain will be successful in lowering the blood pressure in human subjects with high blood pressure levels.

2. Since there are several contributing factors that can affect blood pressure levels, list five factors that will be constant between the experimental and control groups.
Age, smoker or non-smoker, sex, present blood pressure, diet, stress, amount of daily exercise, percent body fat, weight, family history, daily or weekly alcohol consumption, cholesterol level, etc.

3. Based on the factors listed in Question 2, using complete sentences explain why certain criteria need to be used in choosing the participant in this study.
The categories would have to be chosen to match the people in the two different groups as closely as possible to the factors listed in Question #2.

4. Once the list of the participants has been created, using complete sentences explain how they will be selected to be a member of either the experimental or control group.
I would divide up the participants randomly in the control and experimental groups.

5. Using complete sentences, explain what measurements and/or tests will be made on the experimental and control groups to judge the efficiency of Alamain, and how often measurements or test will be taken.
I would check their blood pressure and heart rates at least once a day, once a week, etc. and measure any side effects between the two groups.

6. Using complete sentences, explain what criteria will be used to indicate the success or failure of the drug Alamain to reduce blood pressure levels in humans.
The drug lowered the blood pressure in the experimental group with no harmful side effects.

THE 'BIRD ASSESSMENT'

The College Board (2009) AP® Statistics Free-Response Form B. Question 4 Page 8.
Assessment: [Online
http://apcentral.collegeboard.com/apc/public/repository/ap09_frq_statistics_formb.pdf]
Scoring Guidelines:
http://apcentral.collegeboard.com/apc/public/repository/ap09_statistics_form_b_sgs.pdf
(Used with permission to Nancy Pelaez, npelaez@purdue.edu)

1. Birds have four types of color receptors in the eye. Most mammals have two types of receptors, although primates have three. Birds also have proportionally more nerve connections between the photoreceptors and the brain. Previous research has shown differences between male and female zebrafinches in their tendency to avoid food that has solid colors. Suggest a potential cause for this difference between male and female zebrafinches. Briefly explain.

Because birds have four types of color receptors, they are able to see different wavelengths of light than mammals that have two or three types. The four color receptors also give a broader range of light, possibly allowing the birds to see ultraviolet light. Male zebrafinches are very distinct from female zebrafinches. The males have bright patches of color on their plumage, while females are mostly one solid color. Evolution may have adapted male zebrafinches to be attracted to solid colors so they will easily find a mate. This would explain why males eat solid colored fruit. On the contrary, females may have adapted to be attracted to stripes or patterns of colors. This would explain why females avoid eating solid fruit. Because they avoid solid fruit, one could say they may also avoid other solid females making their chances of mating increase.

2. Good biological knowledge could help you become an entrepreneur. For example, a manufacturer of toxic pesticide granules plans to use a dye to color the pesticide so that birds will avoid eating it. A series of experiments will be designed to find colors or patterns that three bird species (blackbirds, zebrafinches, and geese) will avoid eating. Representative samples of birds will be captured to use in the experiments, and the response variable will be the amount of time a hungry bird will avoid eating food of a

particular color or pattern. a. Previous research has shown that male birds do not avoid solid colors. However, it is possible that males might avoid colors displayed in a pattern, such as stripes. In an effort to prevent males from eating the pesticide, the following two treatments are applied to pesticide granules:
Treatment 1: A red background with narrow blue stripes
Treatment 2: A blue background with narrow red stripes
To increase the power of detecting a difference in the two treatments in the analysis of the experiment, the researcher decided to block on the three species of birds (blackbirds, zebrafinches, and geese). Assuming there are 100 birds of each of the three species, explain how you would assign birds to treatments in such a block design.

Form three blocks based on the species of bird (blackbirds, starlings, and geese) carrying a equal distribution of male: female birds to accomplish the goal of blocking to create groups of homogeneous experimental units. Within each of the three blocks, carry out a completely randomized design by randomly assigning the birds within each block to one of the two treatments. Within block 1, each bird of a particular species (let's say the blackbirds) will be tagged with a unique random number using a random number generator on a calculator, statistical software, or a random number table. The random numbers will be sorted from lowest to highest. The birds with the lowest 50 numbers in the ordered list will receive treatment 1 (red background with narrow blue stripes). The birds with the highest 50 numbers will receive treatment 2 (blue background with narrow red stripes). This method of randomization should be repeated in the other two blocks.

b. What else could the researcher do to increase the power of detecting a difference in the two treatments in the analysis of the experiment? Explain how your approach would increase the power.
To increase power (other than by blocking), the researcher could increase the sample size. This reduces the standard error of the sampling distribution. With a smaller standard error, a test is more likely to be able to detect a difference in results from the two treatments, if such a difference exists.

Typical ‘Evidence of Difficulties’ Examples from RED (Table 2)

Tables SI 1- 3 include response phrases that provide evidence of difficulties that are underlined and coded with a footnote that corresponds to a row in Table 2.

Table SI 1: Typical ‘evidence of difficulties’ from the ‘Shrimp Assessment’ responses.

‘Shrimp Assessment’: A biologist is interested in studying the effect of growth-enhancing nutrients and different salinity (salt) levels in water on the growth of shrimps. The biologist has ordered a large shipment of young tiger shrimps from a supply house for use in the study. The experiment is to be conducted in a laboratory where 10 tiger shrimps are placed randomly into each of 12 similar tanks in a controlled environment. The biologist is planning to use 3 different growth-enhancing nutrients (A. B. and C) and two different salinity levels (low and high).

Student ID	1. List the treatments that the biologist plans to use in this experiment.	2. Using the treatments listed in part (a), describe a completely randomized design that will allow the biologist to compare the shrimps' growth after 3 weeks.	3. Give one statistical advantage to having only tiger shrimps in the experiment. Explain why this is an advantage.	4. Give one statistical disadvantage to having only tiger shrimps in the experiment. Explain why this is a disadvantage.
Anna (Correct)	1. A Low salinity 2. A high salinity 3. B low salinity 4. B high salinity 5. C low salinity 6. C high salinity	A randomized design would be possibly dividing the 6 treatments into each of 12 tanks, so that there are two tanks with each treatment. In order for randomization to occur it might be easiest to use dice and assign each number to its corresponding treatment number. Example: Roll dice 1+ 2; Outcome Die 1= 2 and Die 2= 4. From this you would put treatment two and four in tanks 1 and 2.	The advantage to having only tiger shrimp in the experiment is that you are only using one single species of shrimp. This leads to an advantage because there is less variability within the growth of shrimp. As a result, using only tiger shrimps reduces variance.	One statistical disadvantage to only having only tiger shrimp is that due to the fact we only used one species of shrimp we are not able to make a generalization about all shrimp. Our data only correlates to the experiment performed on tiger shrimps. Therefore we can only make an accurate analysis on this particular species of shrimp.
Beth (Difficulty)	Nutrient A with low salinity, Nutrient B with low salinity, Nutrient C with low salinity, Nutrient A with high salinity, Nutrient B with high salinity, Nutrient C with high salinity, Low salinity with no nutrient, High salinity with no nutrient.[1, 2]	Assign each tank a treatment. Put 12 slips of paper numbered 1-12 in a bowl. With all the shrimp in one tank, one by one randomly assign a shrimp to a tank. Replace the 12 strips to the bowl following each 12 shrimps[3]. By doing this, the biologist is aware of which tanks contain which ingredients but the shrimp are completely randomized.[4]	The tiger shrimps act as the control group[5]. In this, a researcher can confidently expect to find a repetitive response to a given exposure in a group of genetically identical tiger shrimps.[6, 7]	The researcher is only studying the effects of a given ingredient on tiger shrimps. This [doesn't] demonstrate how a given ingredient may affect another type of shrimp.[8] Ultimately it limits the depth of the study.

[1] Area of difficulty 2-f
[2] Area of difficulty 2-c
[3] Area of difficulty 4-h
[4] Area of difficulty 4-f
[5] Area of difficulty 1-a
[6] Area of difficulty 3-e
[7] Area of difficulty 4-a
[8] Area of difficulty 5-c

Table SI 2: Typical 'evidence of difficulties' from the 'Drug Assessment' responses.

'Drug Assessment': The drug ALAMAIN has been developed by the Gentronic Drug Company to lower blood pressure in people whose blood pressure is too high. The drug has been thoroughly tested on animals with positive results. The Gentronic Drug Company feels it is now time for the drug to be tested on humans, and have contacted the Human Improvement Laboratory (HIL) to do the testing. Directions: As chief research scientist at the Human Improvement Laboratory (HIL) you have been assigned the task of developing the human testing program for the new high blood pressure drug Alamain. You and your assistants are to confer on the experimental design of this testing program, and to write a report outlining the program. The report is to be submitted to the chairperson of the HIL Drug-Testing Committee for approval. Complete the following sections as you would include them on your report.

Student ID	1. Using complete sentences state the hypothesis to be tested.	2. Since there are several contributing factors that can affect blood pressure levels, list five factors that will be constant between the experimental and control groups.	3. Based on the factors listed in Question 2, using complete sentences explain why certain criteria need to be used in choosing the participant in this study.	4. Once the list of the participants has been created, using complete sentences explain how they will be selected to be a member of either the experimental or control group.	5. Using complete sentences, explain what measurements and/or tests will be made on the experimental and control groups to judge the efficiency of Alamain, and how often measurements or test will be taken.	6. Using complete sentences, explain what criteria will be used to indicate the success or failure of the drug Alamain to reduce blood pressure levels in humans.
Josh (Correct)	The hypothesis is that the new drug will lower the blood pressure of people with high blood pressure.	They have to be at the same range of high blood pressure, diet, exercise, eating habits, sleep habits, etc.	These factors are important because without a consistency in the individuals chosen we cannot effectively judge how the drug works based on [results for] the control group and the experimental group members.	They will be chosen at random to be part of the experimental or control group. That way they do not have an opinion on how the drug may or may not be helping them.	Blood pressure will be monitored daily and recorded. The progress of people taking the drug will determine its effectiveness.	If people [with high blood pressure], in the experimental group who take the drug consistently have decreased blood pressure, then the drug is effective.
Ken (Difficulty)	We are going to bring in individuals who are willing to test a new drug, Alamain, which we know have only produced good results on animals so far. This drug will be administered to people at low dosages at first[9], and then we will record results and from there calculate the	Hemoglobin levels will remain constant as well as most proteins. The blood vessels will be relaxed and blood will flow smoothly through them because they will expand. [11,12] To lower the pressure we administer hormones that constrict the vessels at a healthy rate. Red blood	Participants cannot be pregnant simply[13] because it will affect the fetus differently than the adult. People older than 35 should not test the drug[14]. These criteria need to be met and not taken lightly because health problems may arise.[15]	The younger, healthier participants will be the experimental group while the not so young will be the control. [16,17]	Experimental groups will receive a couple different dosages to see how each dose affects blood pressure[18], whereas the control will be compared to the experimental to record differences. Measurements can be taken twice daily but no more than that to start for	If the drug does indeed reduce blood pressure, the percentage of those who[se] blood pressure [becomes] normal will be significantly high than that control group.[19, 20]

[9] Area of Difficulty 2-d

Table SI 2: Typical 'evidence of difficulties' from the 'Drug Assessment' responses.

'Drug Assessment': The drug ALAMAIN has been developed by the Gentronic Drug Company to lower blood pressure in people whose blood pressure is too high. The drug has been thoroughly tested on animals with positive results. The Gentronic Drug Company feels it is now time for the drug to be tested on humans, and have contacted the Human Improvement Laboratory (HIL) to do the testing. Directions: As chief research scientist at the Human Improvement Laboratory (HIL) you have been assigned the task of developing the human testing program for the new high blood pressure drug Alamain. You and your assistants are to confer on the experimental design of this testing program, and to write a report outlining the program. The report is to be submitted to the chairperson of the HIL Drug-Testing Committee for approval. Complete the following sections as you would include them on your report.

Student ID	1. Using complete sentences state the hypothesis to be tested.	2. Since there are several contributing factors that can affect blood pressure levels, list five factors that will be constant between the experimental and control groups.	3. Based on the factors listed in Question 2, using complete sentences explain why certain criteria need to be used in choosing the participant in this study.	4. Once the list of the participants has been created, using complete sentences explain how they will be selected to be a member of either the experimental or control group.	5. Using complete sentences, explain what measurements and/or tests will be made on the experimental and control groups to judge the efficiency of Alamain, and how often measurements or test will be taken.	6. Using complete sentences, explain what criteria will be used to indicate the success or failure of the drug Alamain to reduce blood pressure levels in humans.
	correct amount of Alamain that should be given to each person.[10]	cells will remain at the same constant rate and will not be affected.			safety precautions.	

[11] Area of Difficulty 2-g
[12] Area of Difficulty 1-b
[13] Area of Difficulty 1-b
[14] Area of Difficulty 1-b
[15] Area of Difficulty 4-c
[16] Area of Difficulty 1-b
[17] Area of Difficulty 4-d
[18] Area of Difficulty 2-d
[19] Area of Difficulty 3-g
[20] Area of Difficulty 5-c
[10] Area of Difficulty 2-b

Table SI 3: Typical ‘evidence of difficulties’ from the ‘Bird Assessment’ responses.

‘Bird Assessment’: Birds have four types of color receptors in the eye. Most mammals have two types of receptors, although primates have three. Birds also have proportionally more nerve connections between the photoreceptors and the brain. Previous research has shown differences between male and female zebra finches in their tendency to avoid food that has solid colors. A manufacturer of toxic pesticide granules plans to use a dye to color the pesticide so that birds will avoid eating it. A series of experiments will be designed to find colors or patterns that three bird species (blackbirds, zebra finches, and geese) will avoid eating. Representative samples of birds will be captured to use in the experiments, and the response variable will be the amount of time a hungry bird will avoid eating food of a particular color or pattern. Previous research has shown that male birds do not avoid solid colors. However, it is possible that males might avoid colors displayed in a pattern, such as stripes. In an effort to prevent males from eating the pesticide, the following two treatments are applied to pesticide granules: Treatment 1: A red background with narrow blue stripes; Treatment 2: A blue background with narrow red stripes.

Student ID	1. Suggest a potential cause for the difference between male and female zebra finches. Briefly explain.	2. a. To increase the power of detecting a difference in the two treatments in the analysis of the experiment, the researcher decided to block on the three species of birds (blackbirds, zebrafinches, and geese). Assuming there are 100 birds of each of the three species, explain how you would assign birds to treatments in such a block design.	b. What else could the researcher do to increase the power of detecting a difference in the two treatments in the analysis of the experiment? Explain how your approach would increase the power.
Rita (Correct)	These sorts of behavior may have a root in the past. Perhaps at some point early in the zebrafish species' development, food with solid colors had a deleterious effect on the zebrafish's survival and reproduction abilities. If the male now avoids solid color food, there may be a chemical that acts as a spermicide or acts as a testosterone antagonist (blocking testosterone receptors that enable proper reproductive tissues to grow and function properly, and thus leading to a decrease in reproduction rate) or the chemicals in solid foods have some other sort of deleterious effect on the functioning of the zebrafish's body, either through binding to other necessary receptors and blocking them, making the zebrafish hypersensitive to other hormonal signals/neurotransmitters.	Knowing from previous research that male birds do not avoid solid colors, there is no need to test the birds against a control treatment of colorless or solid colored food. As this is the case, each species of bird would be randomly divided into two groups, with one group receiving treatment 1 and the other group receiving treatment 2 (that is, 50 blackbirds would receive treatment 1, 50 blackbirds would receive treatment 2, and likewise for zebrafinches and geese).	Ensuring that all of the birds being tested are as similar as possible except for the treatment is best. This entails that all birds have the same gender, are roughly the same age, come from very similar habitats, and are in overall good health (no underlying conditions such as currently suffering from a given disease). With all of these potential differences eliminated, the birds would be made different in only one respect: their treatment. In this manner, one would be able to confidently declare that differences in the response variable (in this case, the frequency of avoiding or not avoiding food given the particular treatment) can be *[attributed to]* the difference in treatment.

Table SI 3: Typical 'evidence of difficulties' from the 'Bird Assessment' responses.

'Bird Assessment': Birds have four types of color receptors in the eye. Most mammals have two types of receptors, although primates have three. Birds also have proportionally more nerve connections between the photoreceptors and the brain. Previous research has shown differences between male and female zebra finches in their tendency to avoid food that has solid colors. A manufacturer of toxic pesticide granules plans to use a dye to color the pesticide so that birds will avoid eating it. A series of experiments will be designed to find colors or patterns that three bird species (blackbirds, zebra finches, and geese) will avoid eating. Representative samples of birds will be captured to use in the experiments, and the response variable will be the amount of time a hungry bird will avoid eating food of a particular color or pattern. Previous research has shown that male birds do not avoid solid colors. However, it is possible that males might avoid colors displayed in a pattern, such as stripes. In an effort to prevent males from eating the pesticide, the following two treatments are applied to pesticide granules: Treatment 1: A red background with narrow blue stripes; Treatment 2: A blue background with narrow red stripes.

Student ID	1. Suggest a potential cause for the difference between male and female zebra finches. Briefly explain.	2. a. To increase the power of detecting a difference in the two treatments in the analysis of the experiment, the researcher decided to block on the three species of birds (blackbirds, zebrafinches, and geese). Assuming there are 100 birds of each of the three species, explain how you would assign birds to treatments in such a block design.	b. What else could the researcher do to increase the power of detecting a difference in the two treatments in the analysis of the experiment? Explain how your approach would increase the power.
Sara (Difficulty)	The reason for these differences between the two sexes could have to do with the fact that one sex is the main contributor of food to their young[21]. The sex that is feeding the young, which in most cases would be the female passing the food, might want to avoid certain foods that would be harmful to the young. The zebra finch is able to recognize harmful foods, in this case foods with solid colors, and bring back food for their young that will be beneficial to them.	You could set up three separate areas having one species assigned to one of the three[22]. In each of the area you could spread the same amount of each of the two treatments and allow the birds to be in the areas for a set amount of time.[23] After the time is up, you could collect the remaining seeds and see which treatment was eaten the most and which treatment the birds avoided the most[24].	You could repeat the experiment but this time allowing all three of the species to be in the same area. [25,26] Since in reality they all will be in the same area together and not separated, this experiment would take into account any competition that might take place[27]. This would increase the power by determining which seed the birds compete over and which seed the birds ignore[28].

[21] Area of Difficulty 1-c
[22] Area of Difficulty 1-c
[23] Area of Difficulty 4-e
[24] Area of Difficulty 3-g
[25] Area of Difficulty 2-d
[26] Area of Difficulty 2-f
[27] Area of Difficulty 2-g
[28] Area of Difficulty 3-c

Additional Examples from the 'Typical Evidence of Difficulties' list from Table 2

Table SI 4: Examples of additional 'typical evidence of difficulties' according to RED from the 'Shrimp Assessment'

'Shrimp Assessment': A biologist is interested in studying the effect of growth-enhancing nutrients and different salinity (salt) levels in water on the growth of shrimps. The biologist has ordered a large shipment of young tiger shrimps from a supply house for use in the study. The experiment is to be conducted in a laboratory where 10 tiger shrimps are placed randomly into each of 12 similar tanks in a controlled environment. The biologist is planning to use 3 different growth-enhancing nutrients (A. B. and C) and two different salinity levels (low and high).

Student ID	1. List the treatments that the biologist plans to use in this experiment.	2. Using the treatments listed in part (a), describe a completely randomized design that will allow the biologist to compare the shrimps' growth after 3 weeks.	3. Give one statistical advantage to having only tiger shrimps in the experiment. Explain why this is an advantage.	4. Give one statistical disadvantage to having only tiger shrimps in the experiment. Explain why this is a disadvantage.
Ariel	The three different growth-enhancing nutrients (A,B, and C) and two different salinity levels (low and high).	Measure how much the shrimps grow in each one of the tanks with the independent variables in them. One tank would be the control with no salt or nutrients[29]. There would then be tanks with no salt but with nutrient A in one, B in another, and C in the last.[30] Then get three more tanks, all with salt, and place nutrient A in one, B in another, and again C in the last.	Size can be compared knowing that the only factors contributing to the differences in growth are from the independent variables since all the shrimp are alike.	The experiment is limited to the just tiger shrimp. This experiment would not explain whether the nutrients would affect any other shrimp other than tiger shrimp alone.
Brett	The different growth enhancing nutrients would be tested in both high and low salinity conditions, as in A in high salinity, A in low salinity, B in high, etc. Also, there would need to be control samples, where shrimp were not given the nutrients[31] and are in both high and low salinity water.	Assuming the shrimp were fed in the same manner, the easiest way to compare the shrimps' growth would be by comparing their weight. Since 10 shrimp are in each tank, comparing the total shrimp weight will give a better result than comparing individual shrimp weights.	The comparisons of weight will be simpler due to all shrimp being expected to grow similarly barring any outside influences	Tiger shrimp could be unaffected by either salinity changes or the nutrients, implying a certain reaction that can't necessarily be justified

Manipulation of Variables. [29] For the shrimp assessment, Ariel suggests treatment groups with a growth enhancing nutrient and no salinity: *"There would be tanks with no salt but with nutrient A in one, B in another, and C in the last"* which shows an error as independent variables are haphazardly applied, in scenarios when the combined effects of two independent variables are to be tested simultaneously, in this case, combination of salt and nutrients (Table 2, Area of Difficulty 2-e).

[30] Additionally Ariel also shows a difficulty with control groups when proposing treatments, *"One tank would be the control with no salt or nutrients."* Here the error is that the control group does not provide natural behavior conditions because absence of the variable being manipulated (salt or nutrients) in the treatment group, results in conditions unsuitable for the experimental subject as the shrimp won't survive in such conditions (Table 2, Area of Difficulty 2-h).

[31] Brett proposes a control where *"...shrimp were not given the nutrients"* which is inappropriate as the experimental goal is to compare among 3 different growth enhancing nutrients and not whether nutrients are required or not. Hence, the difficulty is control group treatment conditions are inappropriate for the stated hypothesis or experiment goal (Table 2, Area of Difficulty 2-i).

Table SI 5: Examples of additional ‘typical evidence of difficulties’ according to RED from the ‘Drug Assessment’

‘Drug Assessment’: The drug ALAMAIN has been developed by the Gentronic Drug Company to lower blood pressure in people whose blood pressure is too high. The drug has been thoroughly tested on animals with positive results. The Gentronic Drug Company feels it is now time for the drug to be tested on humans, and have contacted the Human Improvement Laboratory (HIL) to do the testing. Directions: As chief research scientist at the Human Improvement Laboratory (HIL) you have been assigned the task of developing the human testing program for the new high blood pressure drug Alamain. You and your assistants are to confer on the experimental design of this testing program, and to write a report outlining the program. The report is to be submitted to the chairperson of the HIL Drug-Testing Committee for approval. Complete the following sections as you would include them on your report.

Student ID	1. Using complete sentences state the hypothesis to be tested.	2. Since there are several contributing factors that can affect blood pressure levels, list five factors that will be constant between the experimental and control groups.	3. Based on the factors listed in Question 2, using complete sentences explain why certain criteria need to be used in choosing the participant in this study.	4. Once the list of the participants has been created, using complete sentences explain how they will be selected to be a member of either the experimental or control group.	5. Using complete sentences, explain what measurements and/or tests will be made on the experimental and control groups to judge the efficiency of Alamain, and how often measurements or test will be taken.	6. Using complete sentences, explain what criteria will be used to indicate the success or failure of the drug Alamain to reduce blood pressure levels in humans.
Cara	The drug is effective on people with high blood pressure.[32]	1.Asleep or awake – usually lower when sleeping / 2.Body position - lying down, sitting or standing / 3.Activity level - from not moving to extreme exertion / 4.Smoking – increases blood pressure / 5.Caffeine – increases blood pressure[33]	If the criteria is different there will be a complete different outcome.	They have to come from same age group.	I would have all of the participants sleep for six hours and take their blood pressure before that I would restrict them from having any alcohol caffeine or tobacco product. Then give them the ALAMAIN. Take their blood pressure every hour and record it.	The blood pressure both systolic and diastolic has come down to 140 and 90 after taking the ALAMAIN.
Doug	The administration of the drug Alamain to a group of patients will cause a significant decrease in blood pressure.	Weight, height, age, ethnicity, gender.	High blood pressure may have several different root causes that require different treatments, limit the effectiveness of a treatment, or even make certain treatment side effects occur.	They would be divided randomly to avoid bias.	Blood pressure would need to be measured over the course of several months as the drug would not be immediately effective and it would need to be seen if the drug remained constantly effective. Initial conditions	The effectiveness in lowering blood pressure, the mildness of the side effects, the length of effectiveness, and how many people can be helped by this drug would be useful

[32] ***Manipulation of Variables.*** Cara’s hypothesis (Table SI5), *“The drug is effective on people with high blood pressure”* only carries a treatment variable in the hypothesis statement but an outcome variable is missing as this statement does not mention *“the drug lowering blood pressure”* as a specific outcome (Table 2, Area of difficulty 2-a).

[33] Cara considers irrelevant variables in her experiments by suggesting that properties like, *“Asleep or awake, body positions”* to be maintained constant across experimental groups (Area of difficulty 2-g).

Table SI 5: Examples of additional ‘typical evidence of difficulties’ according to RED from the ‘Drug Assessment’

‘Drug Assessment’: The drug ALAMAIN has been developed by the Gentronic Drug Company to lower blood pressure in people whose blood pressure is too high. The drug has been thoroughly tested on animals with positive results. The Gentronic Drug Company feels it is now time for the drug to be tested on humans, and have contacted the Human Improvement Laboratory (HIL) to do the testing. Directions: As chief research scientist at the Human Improvement Laboratory (HIL) you have been assigned the task of developing the human testing program for the new high blood pressure drug Alamain. You and your assistants are to confer on the experimental design of this testing program, and to write a report outlining the program. The report is to be submitted to the chairperson of the HIL Drug-Testing Committee for approval. Complete the following sections as you would include them on your report.

					would also have to be measured to compare to later to check for side effects.	criteria in measuring the drug[34].
Emma	Because the drug has been proven to be effective in animals, it will be just as effective in humans.	Five factors that should be constant are age, race, medical history, weight, and diet.	In order to test this drug, participants need to be chosen carefully. Weight should be criteria because an obese person is much more likely to have high blood pressure than a person who is of average weight. Also, the diet of the participants need to be taken into special consideration because the blood pressure of someone who eats foods that are high in fat will be much higher than that of a person who eats low-fat foods.	If all the participants fit the criteria, then they can be randomly chosen to be in either group.	The blood pressure of both groups should be taken every week and the results should be compared so as to determine if there is any change in blood pressure levels.	If the results observed in the human experiment is the same, or similar, to that observed in the animal experiment, then the drug is a success. If the results are completely different, then the drug is a failure.[35]
Frieda	ALAMAIN will safely lower blood pressure in humans and have no harmful results.	Gender, age, race, heart conditions, blood pressure range	If you are going to compare two groups, the background has to be similar/same in order to eliminate other variables that could disrupt the results.	Once a certain race is determined, then random selection would be the best. Volunteers will be asked to join the experiment.	Blood pressure should be measured when resting and when exercising. Then the recovering pressure can be measured. It should also be measured every day to make sure it isn't just short term, but long term recovery.	Long term blood pressure recovery is the best method to make sure the pressure remains low forever and not just when initially taken.[36]

[34] ***Measurement of Outcome.*** Doug’s hypothesis indicates the administration of the drug Alamain is supposed to be for a group of patients and not for a large population. But when asked to suggest determination of success of the drug he states, *“How many people can be helped by this drug…”* which suggests an incoherent relationship between treatment and outcome variable (Area of difficulty 3-a).

[35] As a measure to indicate success of the blood pressure drug, Emma writes, *“If the results observed in the human experiment is the same, or similar, to that observed in the animal experiment, and then the drug is a success. If the results are completely different, then the drug is a failure.”* This shows an error that an outcome variable was not listed for the investigation as we don’t know what the student means by results being *“similar or different”* (Area of difficulty 3-f).

[36] The stated outcome by Frieda is not measurable (Area of difficulty 3-d) as it suggests, *“Long term blood pressure recovery is the best method to make sure the pressure remains low forever and not just when initially taken.”* Measuring blood pressure for a certain fixed time period is a feasible measure but *“remaining low forever”* is not when deciding success of developed drug.

Table SI 5: Examples of additional ‘typical evidence of difficulties’ according to RED from the ‘Drug Assessment’

‘Drug Assessment’: The drug ALAMAIN has been developed by the Gentronic Drug Company to lower blood pressure in people whose blood pressure is too high. The drug has been thoroughly tested on animals with positive results. The Gentronic Drug Company feels it is now time for the drug to be tested on humans, and have contacted the Human Improvement Laboratory (HIL) to do the testing. Directions: As chief research scientist at the Human Improvement Laboratory (HIL) you have been assigned the task of developing the human testing program for the new high blood pressure drug Alamain. You and your assistants are to confer on the experimental design of this testing program, and to write a report outlining the program. The report is to be submitted to the chairperson of the HIL Drug-Testing Committee for approval. Complete the following sections as you would include them on your report.

Gage	The clinical trials of this drug will be successful by lowering patient’s blood pressure.[37]	The person’s blood type, cholesterol levels, genetic information, body type, and pre-existing medical conditions.	The new drug may not work on people with a certain blood type or pre-existing condition that may already alter the blood pressure. The cholesterol may inhibit the workings of the drug. Body type may play a role in how the drug is dispersed within the body. Genetic information may make someone naturally immune to the drug.	Certain blood tests would be run. A thorough medical background check would also be necessary to look for any genetic problems or pre-existing conditions that may negatively affect the drug.	Regular testing of blood coagulation would be taken to measure if the blood gets thinner or thicker.[38] I would also take regular measurements of cholesterol levels and blood pressure.	We would have to prove that patients on Alamain had regular and consistent drops in their blood pressure with minimal to no side effects. This would prove that the drug works in the human body.
Harry	ALAMAIN can lower the blood pressure of humans.	The diet menu, the time and kinds of sporting, the living habits and the age, gender and species of humans of the experimental and control group.	Because in this experiment we just want to check the effect of ALAMAIN on the blood pressure of humans, but the factors listed in Question 2 can also affect experiment results.	We have one control group and one experiment group. Just divide all the participants into these two groups randomly.	Measurement: the blood pressure of participants. / How often: three times a day: in the morning after breakfast, at the noon after lunch and at night before sleep.	Whether others can redo this experiment with other participants later and get the same result.[39]
Ina	The drug will be administered to a large group and variation of human subjects and will yield results that will show lower blood pressure levels.	Nutrition, stress, fitness, medication, and smoking will all be constant in the experimental group.	Nutrition is important to make sure an unhealthy or healthy food intake does not throw off results yielded from testing the drug. / Stress greatly increases blood pressure, this needs to be kept constant in all subjects to allow room to make the same difference. / Fitness should be similar	The control group will be comprised of all identical types of people will similar body types and lifestyles. The experimental group can have more of a variation and will be	Blood pressures will be regulated before each dose of Alamain (possibly once a day) and the data will be compiled and analyzed at the end of the study.	The criteria to determine success or failure will be whether the drug causes a significant negative change in blood pressure of the human test subject.

[37] Gage shows an error in this area because according to his hypothesis, *“The clinical trials of this drug will be successful by lowering patient’s blood pressure”* the treatment and outcome variables are reversed (Area of difficulty 3-b) as this statement implies *“ success of the drug”* being the outcome variable while *“lowering blood pressure”* as the treatment or independent variable. It would be accurate if administration of drug was considered as the treatment variable and lowering of blood pressure as outcome variable.

[38] Gage also considers measurement of outcome variables *(“blood coagulation testing”)* that are irrelevant with his hypothesis (Area of difficulty 3-c).

[39] ***Accounting for variability.*** Harry suggests, *“Whether others can redo this experiment with other participants later and get the same result”* as a measure for indicating drug success which shows a problem with replication because he considers replication as repeating the entire experiment at another time with another group of experimental subjects (Table 2, Area of difficulty 4-g).

Table SI 5: Examples of additional 'typical evidence of difficulties' according to RED from the 'Drug Assessment'

'Drug Assessment': The drug ALAMAIN has been developed by the Gentronic Drug Company to lower blood pressure in people whose blood pressure is too high. The drug has been thoroughly tested on animals with positive results. The Gentronic Drug Company feels it is now time for the drug to be tested on humans, and have contacted the Human Improvement Laboratory (HIL) to do the testing. Directions: As chief research scientist at the Human Improvement Laboratory (HIL) you have been assigned the task of developing the human testing program for the new high blood pressure drug Alamain. You and your assistants are to confer on the experimental design of this testing program, and to write a report outlining the program. The report is to be submitted to the chairperson of the HIL Drug-Testing Committee for approval. Complete the following sections as you would include them on your report.

		throughout the test subjects in order to have similar beginning footing and to give no subject an advantage. / Medications should be kept constant and no participant can be given anything additional to avoid some medication making an unexpected change. / Smoking status needs to be similar to avoid giving anyone a disadvantage.	administered with the drug.[40]

Table SI 6: Examples of additional 'typical evidence of difficulties' according to RED from the 'Bird Assessment'

'Bird Assessment': Birds have four types of color receptors in the eye. Most mammals have two types of receptors, although primates have three. Birds also have proportionally more nerve connections between the photoreceptors and the brain. Previous research has shown differences between male and female zebra finches in their tendency to avoid food that has solid colors. A manufacturer of toxic pesticide granules plans to use a dye to color the pesticide so that birds will avoid eating it. A series of experiments will be designed to find colors or patterns that three bird species (blackbirds, zebra finches, and geese) will avoid eating. Representative samples of birds will be captured to use in the experiments, and the response variable will be the amount of time a hungry bird will avoid eating food of a particular color or pattern. Previous research has shown that male birds do not avoid solid colors. However, it is possible that males might avoid colors displayed in a pattern, such as stripes. In an effort to prevent males from eating the pesticide, the following two treatments are applied to pesticide granules: Treatment 1: A red background with narrow blue stripes; Treatment 2: A blue background with narrow red stripes.

Student ID	1. Suggest a potential cause for the difference between male and female zebra finches. Briefly explain.	2. a. To increase the power of detecting a difference in the two treatments in the analysis of the experiment, the researcher decided to block on the three species of birds (blackbirds, zebrafinches, and geese). Assuming there are 100 birds of each of the three species, explain how you would assign birds to treatments in such a block design.	b. What else could the researcher do to increase the power of detecting a difference in the two treatments in the analysis of the experiment? Explain how your approach would increase the power.

[40] Ina shows errors in explaining participant selection: *"The control group will be comprised of all identical types of people with similar body types and lifestyles. The experimental group can have more of a variation and will be administered with the drug."* This is an error because criteria for selecting experimental subjects for treatment vs. control group are biased (body types identical vs. variable) (Table 2, Area of difficulty 4-b). Other problems with variability are found from Ina's suggestion, *"control group will be comprised of all identical types of people"* which indicates flawed understanding of natural variability within a sample of experimental subjects (Area of difficulty 4-a). She also doesn't consider random assignment of control and experimental group participants (Area of difficulty 4-e).

Table SI 6: Examples of additional 'typical evidence of difficulties' according to RED from the 'Bird Assessment'

'Bird Assessment': Birds have four types of color receptors in the eye. Most mammals have two types of receptors, although primates have three. Birds also have proportionally more nerve connections between the photoreceptors and the brain. Previous research has shown differences between male and female zebra finches in their tendency to avoid food that has solid colors. A manufacturer of toxic pesticide granules plans to use a dye to color the pesticide so that birds will avoid eating it. A series of experiments will be designed to find colors or patterns that three bird species (blackbirds, zebra finches, and geese) will avoid eating. Representative samples of birds will be captured to use in the experiments, and the response variable will be the amount of time a hungry bird will avoid eating food of a particular color or pattern. Previous research has shown that male birds do not avoid solid colors. However, it is possible that males might avoid colors displayed in a pattern, such as stripes. In an effort to prevent males from eating the pesticide, the following two treatments are applied to pesticide granules: Treatment 1: A red background with narrow blue stripes; Treatment 2: A blue background with narrow red stripes.

Student ID	1. Suggest a potential cause for the difference between male and female zebra finches. Briefly explain.	2. a. To increase the power of detecting a difference in the two treatments in the analysis of the experiment, the researcher decided to block on the three species of birds (blackbirds, zebrafinches, and geese). Assuming there are 100 birds of each of the three species, explain how you would assign birds to treatments in such a block design.	b. What else could the researcher do to increase the power of detecting a difference in the two treatments in the analysis of the experiment? Explain how your approach would increase the power.
Jack	A potential cause for male and female Zebra Finches difference's in avoiding food that has solid colors could be the result of females needing a certain protein that are found in certain solid or non-solid foods. This may be important in the development of healthy chicks. The males may eat certain solid or non-solid foods in order for the coloration on their feathers to show up brighter. For example, Flamingos eat shrimp that cause the pink coloration of their feathers. It could also hold true for the male Zebra Finch, in order to help attract a mate.	For treatment one, the researcher should test fifty male birds of each species to understand which species of male will avoid a red background with narrow blue stripes. Treatment two will have the remaining fifty male birds of each species in order to understand which species avoids a blue background with narrow red strips. Each species will be tested separately of each other.	The researcher could test different size objects and shapes with either a red background with narrow blue stripes or a blue background [41]with narrow red stripes. This would help the researchers in determining which granules need to be patterned if they know the size of the birds feed. The researcher can also use different colors for testing, such as orange and blue or orange and red. Testing different colors may allow the manufacturer to use more than one patterning of colors or enable them to use the cheaper color that would be used in the dye. It is also a good idea because one or none of the species of birds will avoid seeds in either treatment.

[41] ***Measurement of outcome.*** We found an example of a response by Jack elucidating a difficulty with this area because he suggests to increase the power of detecting a difference in treatments as: *"…test different size objects and shapes with either a red background with narrow blue stripes or a blue background"* This indicates Jack proposes outcome variables (like "size, shapes, variable patterning, price of color") that are irrelevant for his proposed experimental context or provided treatments ("testing how long a bird will avoid colors displayed in stripes") (Table 2, Area of difficulty 3-c).

Inter-rater Reliability Results

10 responses were coded for each assessment. Steps followed for inter-rater reliability exercise are:

- Detailed explanation of rubric in terms of propositional statements for each category, concepts associated with each category and corresponding errors descriptions.
- Explanation of scoring protocol.
- One example for each assessment coded together as an example.
- Raters separated and coded individually.
- Get back together and discuss coding.
- Discuss queries/areas that need clarifications, if any.
- Determine Cohen's kappa values for each area.

Cohen's *kappa* is calculated using the formula $kappa = \frac{f_0\text{-}f_c}{N\text{- }f_c}$ where f_0 denotes the number of responses coded similarly, f_c denotes number of responses that would be expected to be coded the same way by chance alone, and N is the number of units coded by either coder (i.e., if two coders code 50 responses each, N = 50). We calculated *kappa* values for 10 responses from each assessment and compared agreement for 5 major areas. For example, table 1 represents the coding results for the 'Shrimp Assessment'.

Table SI 7: Frequency of Correct vs. Difficulty for 'Shrimp Assessment' by raters A and B

'Shrimp Assessment'		Rater B		Rater A total
		Correct	Difficulty	
Rater A	Correct	15	0	15
	Difficulty	1	31	32
	Rater B total	16	31	47

Number of areas coded as 'correct' by both raters A and B are 15 and number of areas coded as 'difficulty' by rater A but coded 'correct' by rater B is 1. Similarly coded areas by both raters are tallied in the diagonal of the table.

Frequency of areas coded similarly, $f_{0,}$ was 46 (97.87% of codes). Frequency of areas expected to be coded similarly by chance, f_c is calculated using formula:

$$fc = \frac{\text{Rater A}_{\text{correct total}} * \text{Rater B}_{\text{correct total}} / \text{Grand Total} + \text{Rater A}_{\text{difficulty total}} * \text{Rater B}_{\text{difficulty total}} / \text{Grand Total}}{\text{Grant Total}}$$

Thus, fc=$\frac{\frac{15*16}{47}+\frac{32*31}{47}}{47}$=0.56 or 56%. This means f_c is 56% of 46 (frequency of codes coded similarly) is 26.2. Thus inserting these values into the formula for

$kappa = \frac{f_0\text{-}f_c}{N\text{- }f_c} = \frac{46\text{-}26.2}{47\text{-}26.2} = 0.952.$

Interrater reliability was established over 50 RED areas [10 (responses) x 5 (areas)] but for kappa calculations we consider only 47 because 3 areas were classified under 'lack of evidence' and we calculated *kappa* values only for areas coded as 'correct' and 'difficulty'.

Apply the same calculations, *kappa* values for the 'Drug' and the 'Bird Assessment' was found to be 0.929 and 0.896 respectively as shown below.

Table SI 8: Frequency of Correct vs. Difficulty for 'Drug Assessment' by raters A and B

'Drug Assessment'		Rater B		Rater A total
		Correct	Difficulty	
Rater A	Correct	10	0	10
	Difficulty	1	44	45
	Rater B total	11	44	55

Number of observed agreements: 54 (98.18% of the observations). Number of agreements expected by chance: 38.0 (69.09% of the observations). Kappa= 0.929.

Table SI 9: Frequency of Correct vs. Difficulty for the 'Bird Assessment' by raters A and B

'Bird Assessment'		Rater B		Rater A total
		Correct	Difficulty	
Rater A	Correct	13	1	14
	Difficulty	2	36	38
	Rater B total	15	37	52

Number of observed agreements: 49 (94.23% of the observations). Number of agreements expected by chance: 31.1 (59.76% of the observations). Kappa= 0.857

Table SI 10: Frequency of 'correct' and 'difficulty' experimental design areas as measured by three assessments pre (beginning) and post (after) semester.

Areas of Experimental Design Difficulty	**'Shrimp Assessment'**	**Pre (spring 2010; n =40[a])**	**Post (spring 2009; n =40[b])**	**p-value[c] from Fisher's test**	**Interrater Agreement[d] (Cohen's *kappa*)**
Variable Property of an Experimental Subject	Correct	19	31	0.019**	0.90[+]
	Difficulty	18	9		
Manipulation of Variables	Correct	4	17	0.008***	
	Difficulty	27	22		
Measurement of Outcome	Correct	11	24	0.114	
	Difficulty	9	6		
Accounting for Variability	Correct	3	11	0.040**	
	Difficulty	33	29		
Scope of Inference	Correct	2	13	0.004***	
	Difficulty	32	26		

Areas of Experimental Design Difficulty	**'Drug Assessment'**	**Pre (spring 2012; n =31[a])**	**Post (spring 2011; n =40[b])**	**p-value[c] from fisher's test**	**Interrater Agreement[d] (Cohen's *kappa*)**
Variable Property of an Experimental Subject	Correct	13	31	0.003***	0.94[+]
	Difficulty	18	9		
Manipulation of Variables	Correct	4	13	0.092*	
	Difficulty	26	27		
Measurement of Outcome	Correct	8	25	0.007***	
	Difficulty	21	15		
Accounting for Variability	Correct	8	18	0.134	
	Difficulty	22	21		
Scope of Inference	Correct	2	9	0.096*	
	Difficulty	28	29		

Table SI 10 ***continued***

Areas of Experimental Design Difficulty	**'Bird Assessment'**	**Pre (spring 2011; n =40[a])**	**Post (spring 2010; n =40[b])**	**p-value[c] from fisher's test**	**Interrater Agreement[d] (Cohen's *kappa*)**
Variable Property of An Experimental Subject	Correct	12	16	0.482	0.86^{+}
	Difficulty	27	24		
Manipulation of Variables	Correct	4	14	0.015**	
	Difficulty	35	26		
Measurement of Outcome	Correct	9	16	0.025**	
	Difficulty	18	8		
Accounting for Variability	Correct	4	7	0.516	
	Difficulty	34	31		
Scope of Inference	Correct	2	6	0.264	
	Difficulty	33	32		

[ab] Categories where frequency for correct and difficulty is less than the total *n* indicates that remaining responses were classified under 'Lack of Evidence' in those cases.
[c] $p<0.01$ = ***; $p<0.05$**; $p<0.1$ =*
[d] According to Landis and Koch (1977) a *kappa* value >0.70^{+} indicates a high degree of interrater agreement .

Table SI 11: Pre and post % differences in 'correct', 'difficulty' and 'lack of evidence' for five areas of experimental design knowledge

'Shrimp Assessment'	**Variable property of an experimental subject (%)**	**Manipulation of Variables (%)**	**Measurement of Outcome (%)**	**Accounting for Variability (%)**	**Scope of Inference (%)**
Correct	29.5	32.5	32.5	20	27.5
LOE	-8	-20	-25	-10	-12.5
Difficulty	-22.5	-12.5	-7.5	-10	-15
'Drug Assessment'	**Variable property of an experimental subject (%)**	**Manipulation of Variables (%)**	**Measurement of Outcome (%)**	**Accounting for Variability (%)**	**Scope of Inference (%)**
Correct	35.56	19.60	36.69	19.19	16.05
LOE	0.00	-3.23	-6.45	-0.73	1.77
Difficulty	-35.56	-16.37	-30.24	-18.47	-17.82
'Bird Assessment'	**Variable property of an experimental subject (%)**	**Manipulation of Variables (%)**	**Measurement of Outcome (%)**	**Accounting for Variability (%)**	**Scope of Inference (%)**
Correct	10	25	17.5	7.5	10
LOE	-2.5	-2.5	7.5	0	-7.5
Difficulty	-7.5	-22.5	-25	-7.5	-2.5

GLOSSARY OF TERMS (in alphabetical order)

Control: An experimental baseline against which an effect of the treatment conditions may be compared (Holmes, Moody & Dine, 2011).

Control group: the "untreated" group with which an experimental group (or treatment group) is contrasted. It consists of units of study that did not receive the treatment whose effect is under investigation (Gill & Walsh, 2010).

Correlation relationship: Two variables are said to be correlated if an observed change in the level of one variable is accompanied by a change in the level of another variable. The change may be in the same direction (positive correlation) or in the opposite direction (negative correlation). Note that correlation does not imply causality. It is possible for two variables to be associated with each other without one of them causing the observed behavior in the other. When this is the case it is usually because there is a third (possibly unknown) causal factor (NIST/SEMATECH, 2003)

Cause and effect relationship: There is a causal and effect relationship between two variables if a change in the level of one variable (independent variable) causes an effect in the other variable (dependent variable). To establish a cause and effect relationship, one must gather the data by experimental means, controlling unrelated variables which might confound the results. Having gathered the data in this fashion, if one can establish that the experimentally manipulated variable is correlated with the dependent variable, then one should be (somewhat) comfortable in making a causal inference. That is, when the data have been gathered by experimental means and confounds have been eliminated, correlation does imply causation (NIST/SEMATECH, 2003; Wuensch, 2001).

Factors: the specific treatments or experimental conditions (the independent variables) (Dasgupta et al., 2013).

Hypothesis: A testable statement that carries a predicted association between a treatment and outcome variable. An investigator designs an experiment to test the hypothesis, and the experimental results are used to evaluate the hypothesis for confirmation or refutation (Ruxton & Colegrave, 2006).

Outcome (dependent) variable: A factor under investigation where it is reasonable to aruge that there may be a relationship with an independent variable. The dependant variable is measurable in terms of units. (Holmes, Moody & Dine, 2011).

Outside/unrelated/control/confounding variables: Any factors (s) that may influence your observations/experiment but is not the factor you are investigating. (Holmes, Moody & Dine, 2011).

Population: All individuals of a defined group appropriate for collecting information for a particular investigation goal (Dasgupta et al., 2013).

Random (representative) sample: A sample where all experimental subjects from a target demographic have an equal chance of being selected in the control or treatment group. An appropriate representative sample size is one that averages out any variations not controlled for in the experimental design (The College Board, 2006).

Randomization: A random sample is selected from a target population; units are then assigned to different treatment groups (Ramsey & Schafer, 2002).

Replication: Replication is performed to assess natural variability, by repeating the same manipulations to several experimental subjects (or units carrying multiple subjects), as appropriate under the same treatment conditions (Quinn & Keough, 2002).

Sample: A random (smaller) group of representative individuals selected from the population, from which data is collected and conclusions are drawn about the population (Dasgupta et al., 2013)

Subject: The individuals to whom the specific variable treatment or experimental condition is applied. Each experimental subject carries a variable property (Dasgupta et al., 2013).

Treatment (independent) variable: The factor (s) in your experiment whose effect you are examining (Holmes, Moody & Dine, 2011)

Treatment group: A group of experimental subjects or units that are exposed to experimental conditions varying in a specific way (Dasgupta et al., 2013)

Unit: The group of individuals to which the specific variable treatment or experimental condition is applied (Dasgupta et al., 2013)

Variable: A certain property of an experimental subject that can be measured and that has more than one condition (Dasgupta et al., 2013).

Variation: when observations within your data set do not all have the same value (Holmes, Moody & Dine, 2011).

Variability: sources of variability in the experimental design of biological study are often divided into two categories: biological variability (variability due to subjects, organisms, and biological samples) and technical variability (variability due measurement, instrumentation, and sample preparation) (Box et al. 2005; Cox and Reid 2000).